Basic Statistics

#### **Descriptive Analytics for Numerical Columns**:

*Step 1:*

Importing the dataset into Juypter Notebook by using the library Pandas:

* import pandas as pd
* df = pd.read\_csv("sales\_data\_with\_discounts.csv")

*Step 2:*

Identifying numerical columns in the dataset.

* df.info()

Volume, Avg Price , Total Sales Value , Discount Rate ,Discount Amount , Net Sales Value these are the numerical columns in the dataset.

*Step 3:*

Calculating the mean, median, mode, and standard deviation for these numerical columns.

**Syntax to calculate mean ,median, mode and standard deviation**:

* print("mean of the Column name:",df['Column name'].mean())
* print("median of the 'Column name:",df[''Column name '].median())
* print("mode of the 'Column name:",df[''Column name '].mode())
* print("standard deviation of the 'Column name:",df['Column name '].std())

**Values for Volume**:

mean of the volume: 5.066666666666666

median of the volume: 4.0

mode of the volume: 0 3

standard deviation of the volume: 4.231602391213926

**Values for Avg Price**:

mean of the Avg Price: 10453.433333333332

median of the Avg Price: 1450.0

mode of the Avg Price: 0 400

1 450

2 500

3 1300

4 8100

standard deviation of the Avg Price: 18079.90483993645

**Values for Discount Rate:**

mean of the DR: 15.15524189526222

median of the DR: 16.57776565

mode of the DR: 0 5.007822

1 5.055218

2 5.059801

3 5.072124

4 5.084108

...

445 19.940610

446 19.965340

447 19.976879

448 19.982257

449 19.992407

standard deviation of the DR: 4.220602153266692

**Values for Discount Amount:**

mean of the Discount Amount: 3346.4994243526

median of the Discount Amount: 988.93373325

mode of the Discount Amount: 0 69.177942

1 73.025199

2 93.649161

3 94.682735

4 102.705757

...

445 17900.983730

446 21153.498820

447 21496.675370

448 25328.224200

449 25738.022190

standard deviation of the Discount Amount: 4509.902963308047

**Values for Net Sales Value:**

mean of the Net Sales Value: 30466.33613109089

median of the Net Sales Value: 4677.7880595

mode of the Net Sales Value: 0 326.974801

1 330.822058

2 466.350839

3 485.317265

4 496.607104

...

445 162603.186400

446 163915.971000

447 163967.080100

448 166263.161300

449 179507.479000

standard deviation of the Net Sales Value: 46358.65662389911

**Values for Total Sales Value:**

mean of the Total Sales Value: 33812.83555555555

median of the Total Sales Value: 5700.0

mode of the Total Sales Value: 0 2430

standard deviation of the Total Sales Value: 50535.07417255328

**Brief interpretation:**

### **Mean:**

The mean, often referred to as the average, is the sum of all the values in a data set divided by the number of values. It provides a central value of the data set.

### **Median:**

The median is the middle value of a data set when the values are arranged in ascending or descending order. If there is an even number of values, the median is the average of the two middle values. It represents the 50th percentile of the data.

### **Mode:**

The mode is the value that appears most frequently in a data set. There can be more than one mode if multiple values have the same highest frequency, or there may be no mode if all values are unique.

### **Standard Deviation:**

The standard deviation measures the amount of variation or dispersion in a data set. It indicates how much the values in the data set deviate from the mean on average. A low standard deviation means the values are close to the mean, while a high standard deviation indicates the values are spread out over a wider range.

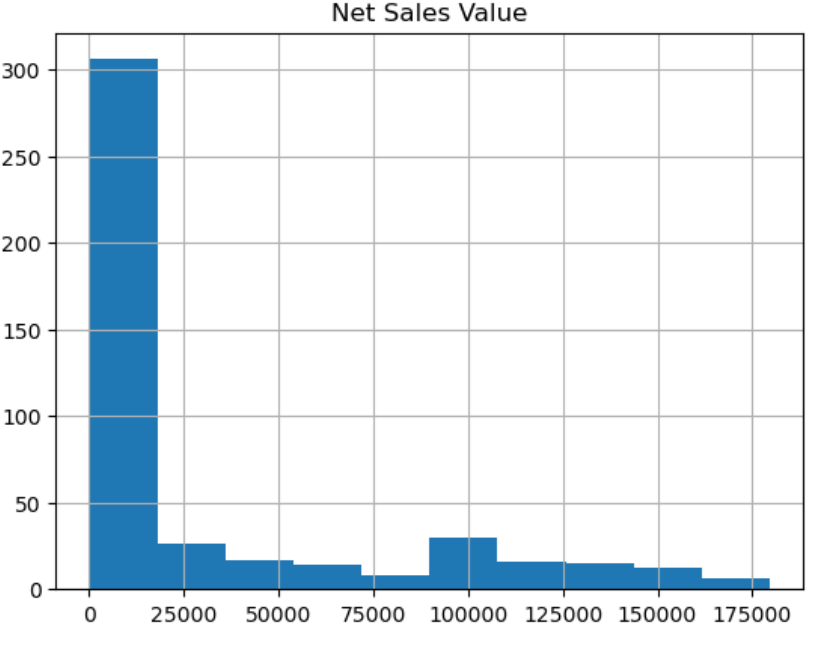
#### **Data Visualization**:

**Plotting Histogram for the numerical column:**

Volume, Avg Price , Total Sales Value , Discount Rate ,Discount Amount , Net Sales Value these are the numerical columns in the dataset.

**Syntax for Plotting histogram and calculating Skewness and Kurtosis:**

* df1.hist('Column name')
* print("Skewness of Column name:",df1['Column name'].skew())
* print("Kurtosis of Column name:",df1['Column name'].kurt())



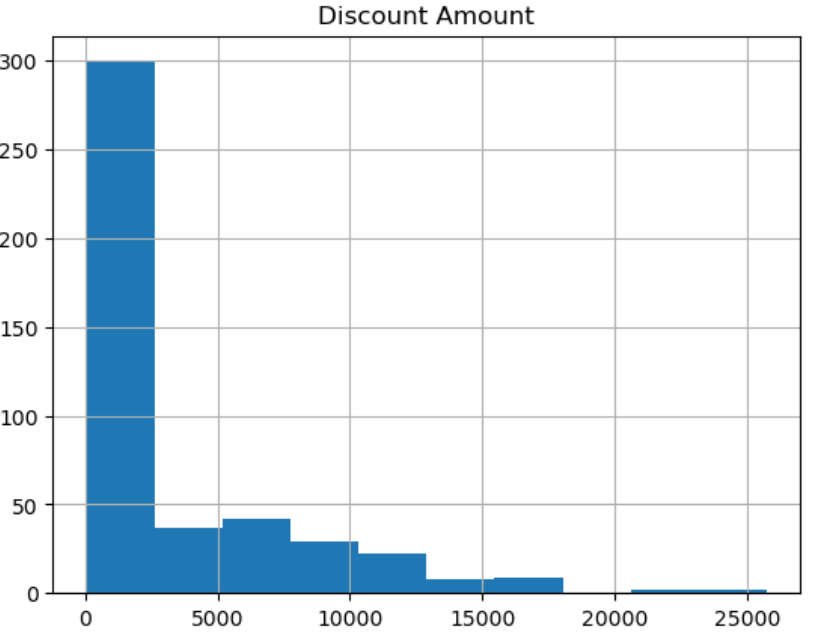
**Observation:**

**Skewness:**

Here the skewness of 1.54 is indicating that the distribution is positively skewed. Positively skewed distributions are having a longer tail on the right side than compared to the left side. Also it is suggesting that there are more extreme values on the right side of the distribution.

**Kurtosis:**

Here the kurtosis value is 1.012 suggesting that the distribution is mesokurtic. Mesokurtic distributions have kurtosis close to that of a normal distribution (where kurtosis = 3). In other words the distribution contains tails and a peak similar to normal distribution Mesokurtic distributions contains a bell-shaped curve that is similar to a normal distribution.



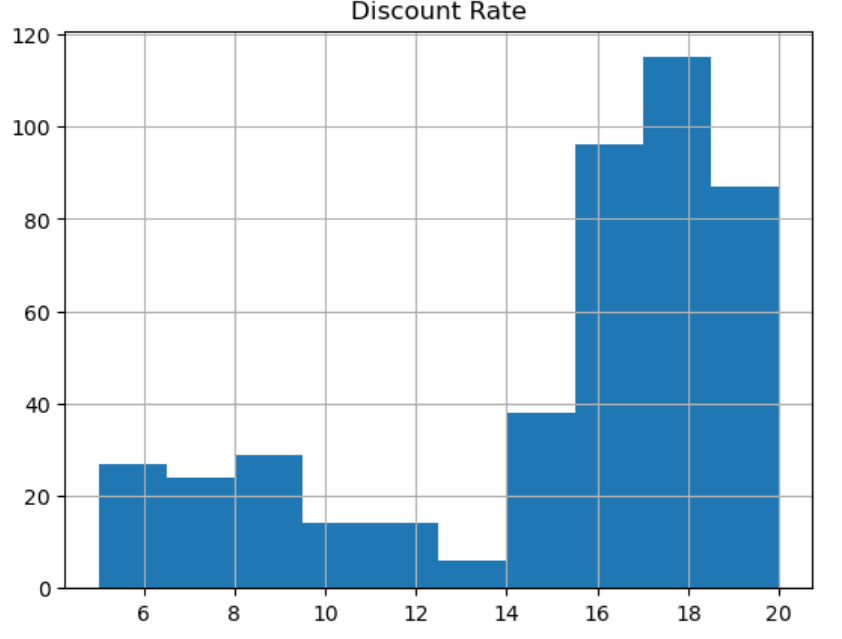
**Observation:**

**Skewness**:

Here the skewness is 1.913 means the distribution is positively skewed. In a positively skewed distribution, the tail will be on the right side (the side which it contains larger data values) is longer than the other side (left side). In other words it is suggesting that there are more extreme values on the right side of the distribution.

**Kurtosis:**

The kurtosis value is 3.83 indicating that the distribution is leptokurtic. Leptokurtic distributions will be having heavier tails and a higher peak than compared to a normal distribution. In other words, there is an increment of extreme values or outliers in the distribution, and in resulting of a sharp and heavier tails.



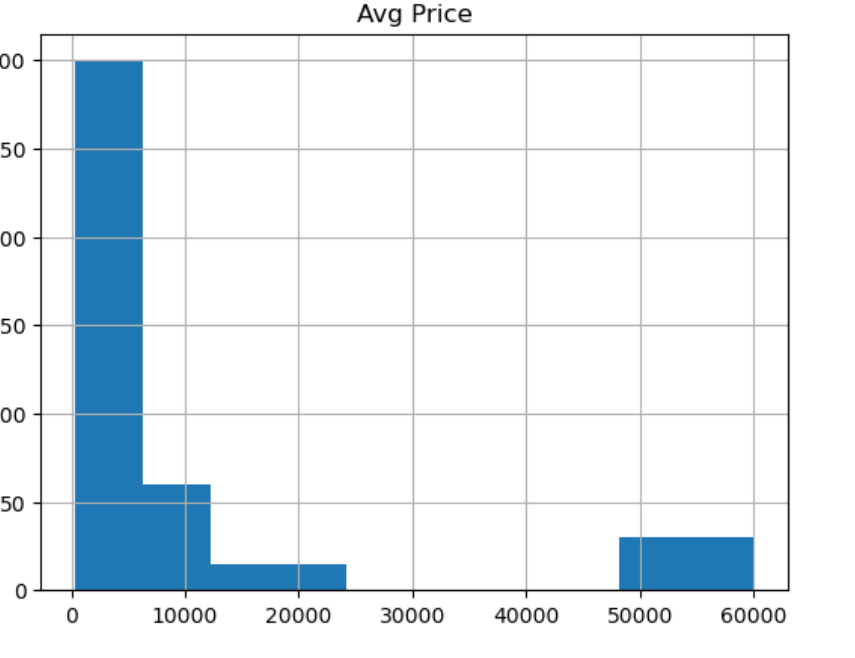
**Observation:**

**Skewness:**

Here the skewness is -1.0623 is indicating a strong negative skewness in the data distribution. The distribution is significantly skewed towards left, in other words it is having a longer tail on the left side and is denser towards the right side.

**Kurtosis:**

Here the kurtosis value obtained is -0.1785 and it is suggesting that the distribution is platykurtic, means it is having thinner tails than compared a normal distribution. In other words, the tails of the distribution are lighter than the tails of a normal distribution .Also indicates that the distribution has lesser extreme values or outliers than a normal distribution.



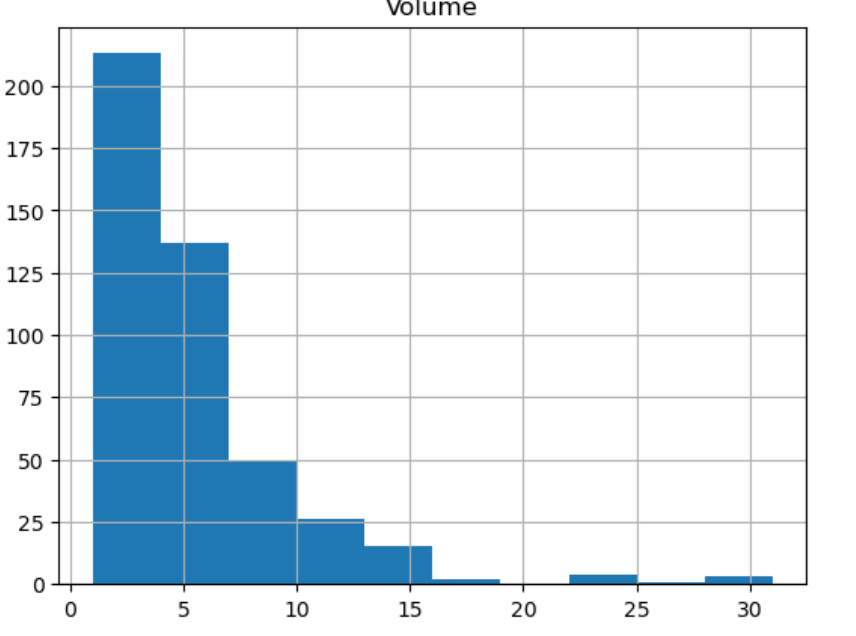
**Observation:**

**Skewness:**

Here the skewness is 1.908 it indicates that the distribution of the data is positively skewed, means the tail on the right side of the distribution is longer than the left side. In other words we can infer that there are more extreme values on the right side of the distribution compared to the left side.

**Kurtosis:**

Here the kurtosis value is 2.07565 it is indicating that the distribution of the data has a higher peak and heavier tails than a normal distribution (k=3). This means that the data is a leptokurtic curve in which , it is having more extreme values and also more peaked than a normal distribution.



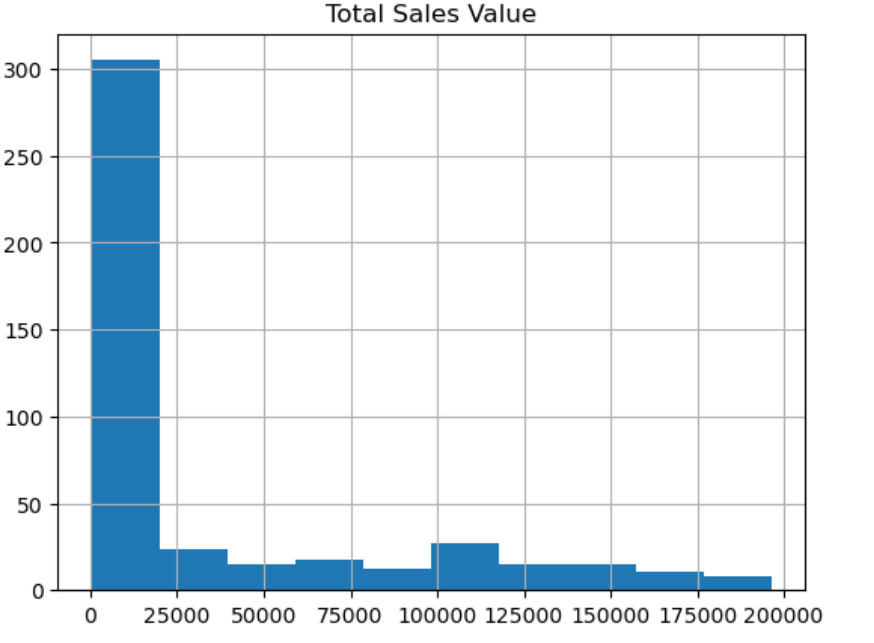
**Observation:**

**Skewness:**

Skewness indicates the asymmetry of the distribution in a data set. Here skewness is 2.73, we can get the inference that distribution is positively skewed. In a positively skewed distribution .

**Kurtosis:**

In general it describes the shape of the probability distribution curve. Here the kurtosis value is 10.258, since the value is very high it indicates that there are more outliers that is more number of tails.



**Observation:**

**Skewness:**

Skewness value of 1.53 indicates that it is will be a positive skewness in the data distribution. The distribution is skewed towards the right, means it is having a longer tail on the right side and is concentrated towards the left side. The skewness is suggesting that the presence of outliers or extreme values on the right side of the distribution.

**Kurtosis:**

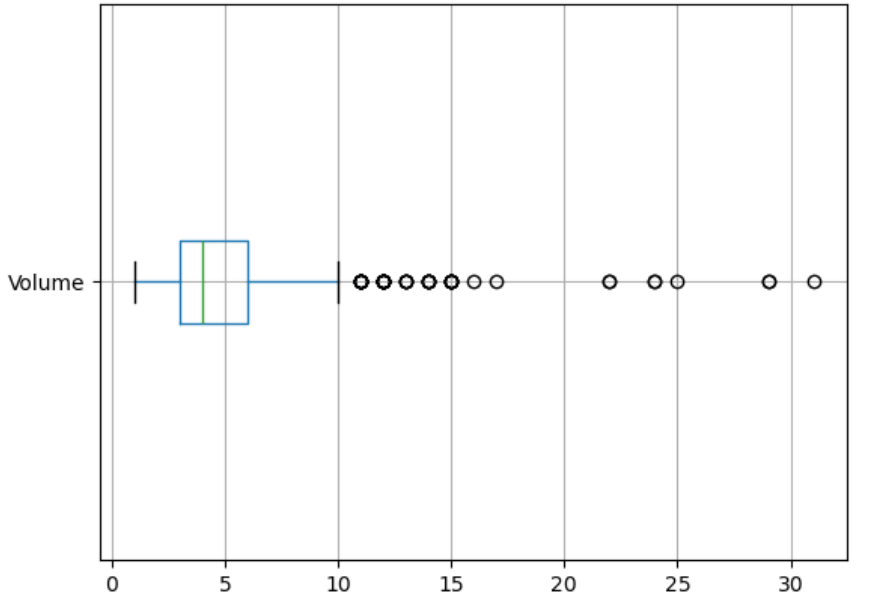
Here the kurtosis value is 1.025 and it is suggesting that the distribution of the data is close to a normal distribution, or mesokurtic. In other words, the distribution is similar to a normal distribution in terms of its peakedness and the thickness. We can conclude that it's neither particularly peaked (leptokurtic)nor particularly flat (platykurtic).

**Plotting Box Plot for the numerical columns** **and observing outliers and inter quartile ranges:**

Volume, Avg Price , Total Sales Value , Discount Rate ,Discount Amount , Net Sales Value these are the numerical columns in the dataset.

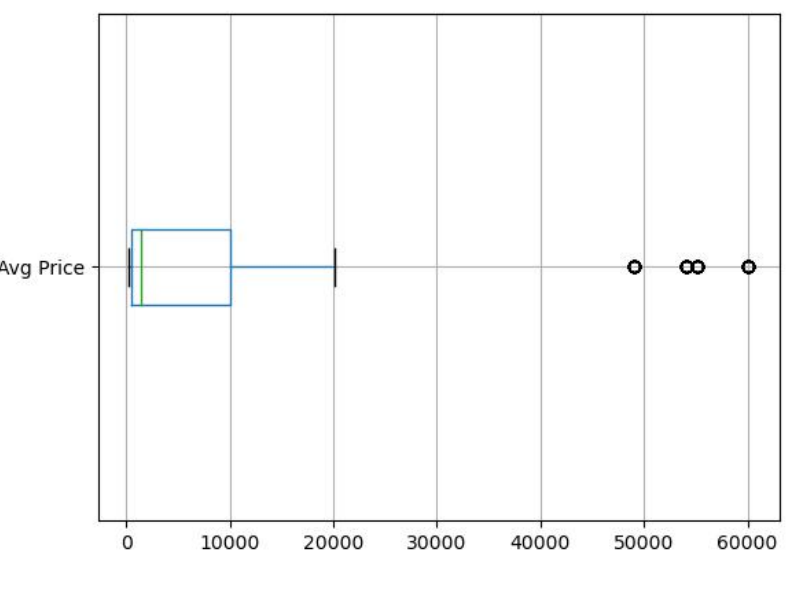
**Syntax for Plotting Box Plot:**

* df1.boxplot(column = 'Column name' ,vert=False)



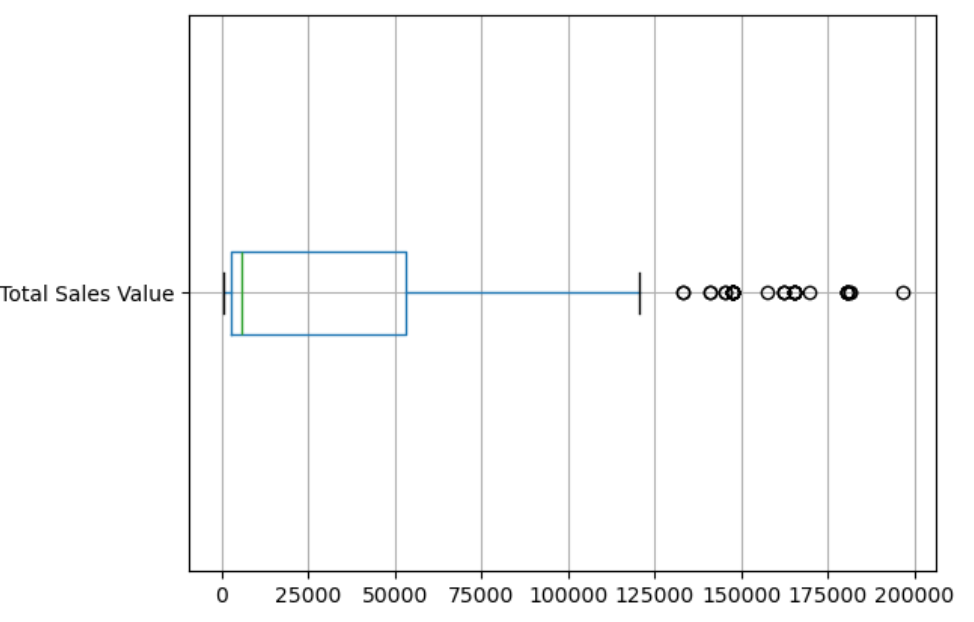
**Observations:**

* Most of the data points lies on the left side.
* The box plot is present over left side therefore the assumption is most of the data points lies on left side and thus it is positively skewed.
* 25th percentile is 3.0.
* 75th percentile that is Q3=6.0
* Inter quartile range is 3.0.
* Outliers are lying above upper whisker length.



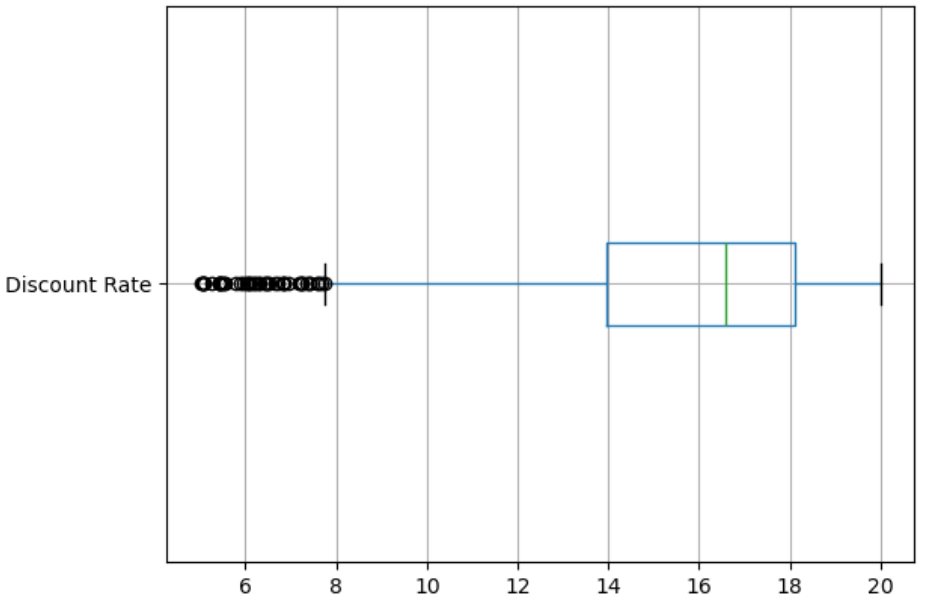
**Observations:**

* Most of the data points lies on the left side.
* The box plot is present over left side therefore the assumption is most of the data points lies on left side and thus it is positively skewed.
* 25th percentile is 465.
* 75th percentile that is Q3=10100
* Inter quartile range is 9635.0.
* Outliers are lying above upper whisker length.



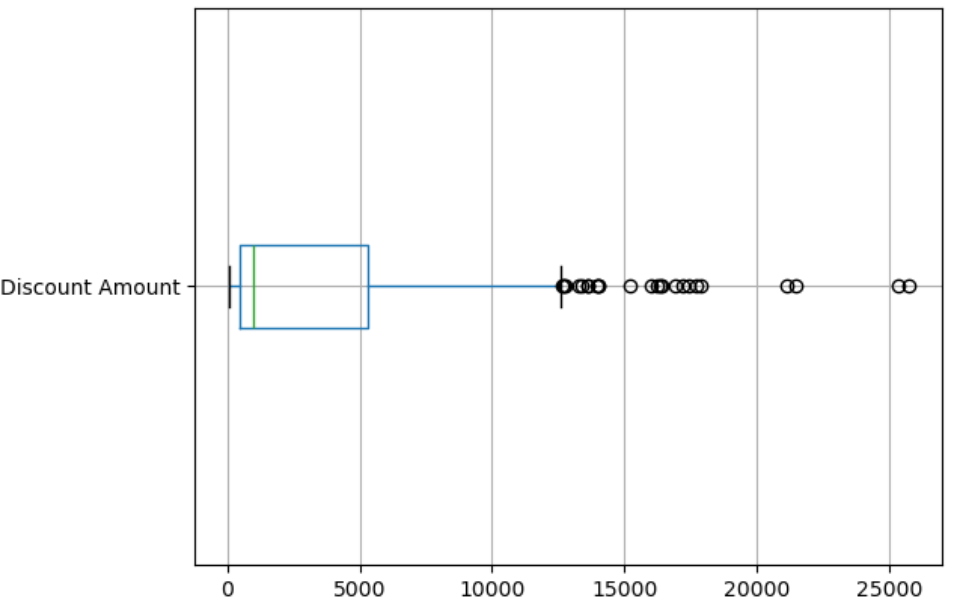
**Observations:**

* Most of the data points lies on the left side.
* The box plot is present over left side therefore the assumption is most of the data points lies on left side and thus it is positively skewed.
* 25th percentile is 2700.
* 75th percentile that is Q3=53200
* Inter quartile range is 50500.0.
* Outliers are lying above upper whisker length



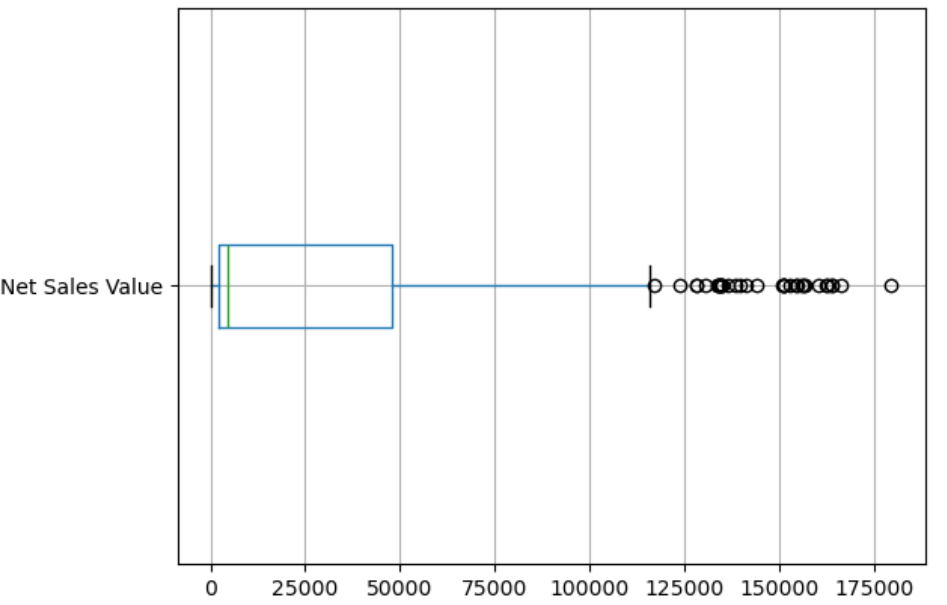
**Observations:**

* Most of the data points lies on the right side.
* The box plot is present over right side therefore the assumption is most of the data points lies on right side and thus it is negatively skewed.
* 25th percentile is 13.965.
* 75th percentile that is Q3=18.115.
* Inter quartile range is 4.15.
* Outliers are lying less than(before) lower whisker length



**Observations:**

* Most of the data points lies on the left side.
* The box plot is present over left side therefore the assumption is most of the data points lies on left side and thus it is positively skewed.
* 25th percentile is 460.5.
* 75th percentile that is Q3=5316.5.
* Inter quartile range is 4856.04.
* Outliers are lying above upper whisker length.



**Observations:**

* Most of the data points lies on the left side.
* The box plot is present over left side therefore the assumption is most of the data points lies on left side and thus it is positively skewed.
* 25th percentile is 2202.21
* 75th percentile that is Q3=47847.9
* Inter quartile range is 45645.7.
* Outliers are lying above upper whisker length

**Bar Chart Analysis for Categorical Column:**

***Step 1:***

* + Identifying Categorical columns in the dataset are :SKU, BU, Brand, Model.

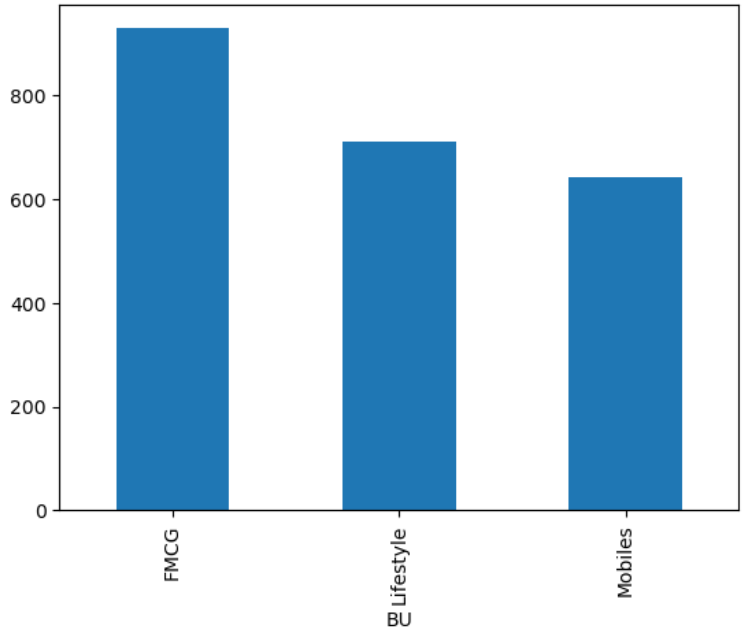
***Step 2:*** Creating bar charts to visualize the frequency or count of each category.

Syntax for counting and creating bar chat:

* x1 = df.groupby('Categorical Column')['Numerical Column'].sum()
* x1
* x1.plot(kind = 'bar')

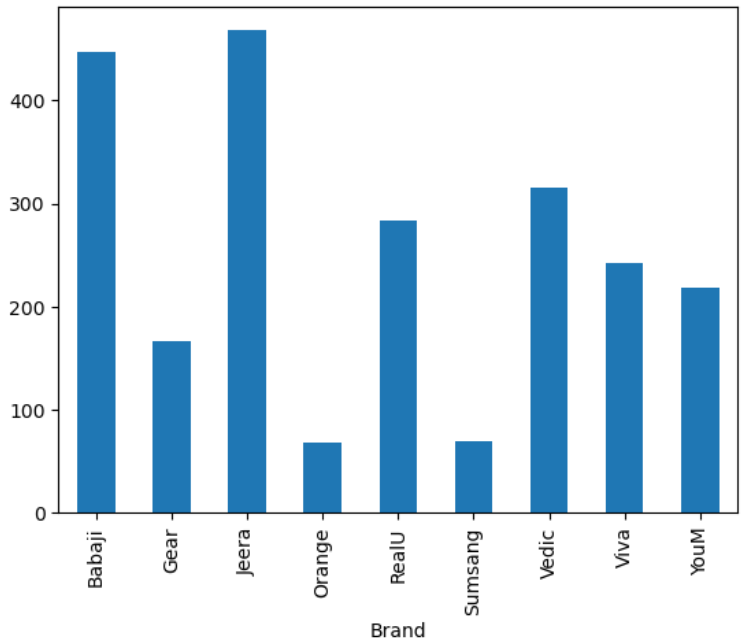
**Performing Bar Graph of “Volume” with categorical columns:**

* “**Volume” Numerical Column with ‘BU’ Categorical Column**.



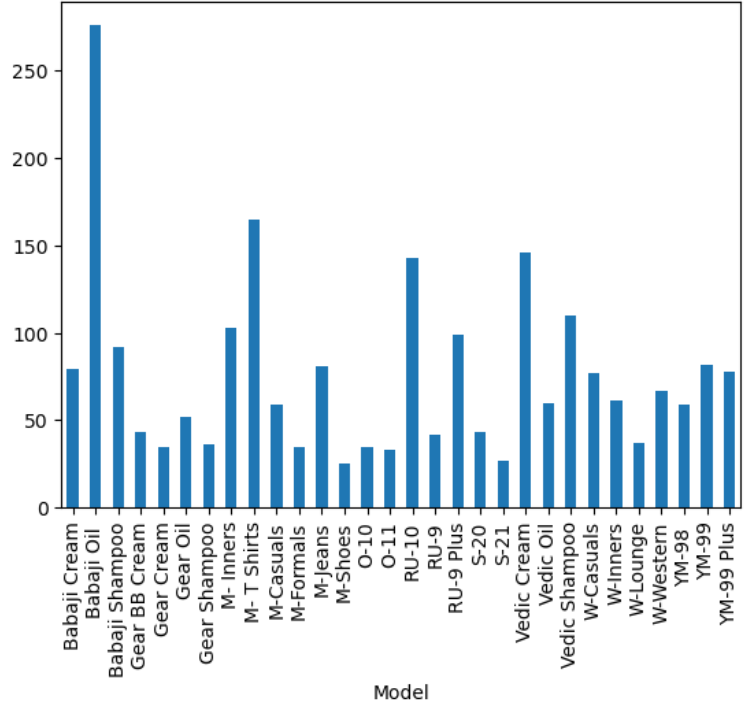
**Insights:**

* Highest frequency columns are FMCG.
* Lowest frequency columns are Mobiles
* “**Volume” Numerical Column with ‘Brand’ Categorical Column.**



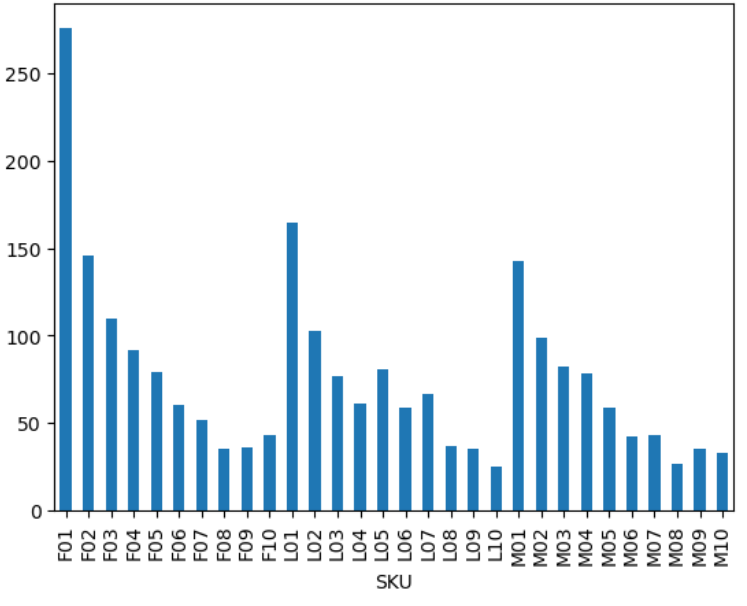
**Insights**:

* Highest frequency columns are Babaji and Jeera.
* Lowest frequency columns are Orange and Samsung
* **“Volume” Numerical Column with ‘Model’ Categorical Column.**



**Insights:**

* Highest frequency columns are Babaji Oil.
* Lowest frequency columns are M-Shoes and S-21.
* **“Volume” Numerical Column with ‘SKU’ Categorical Column.**

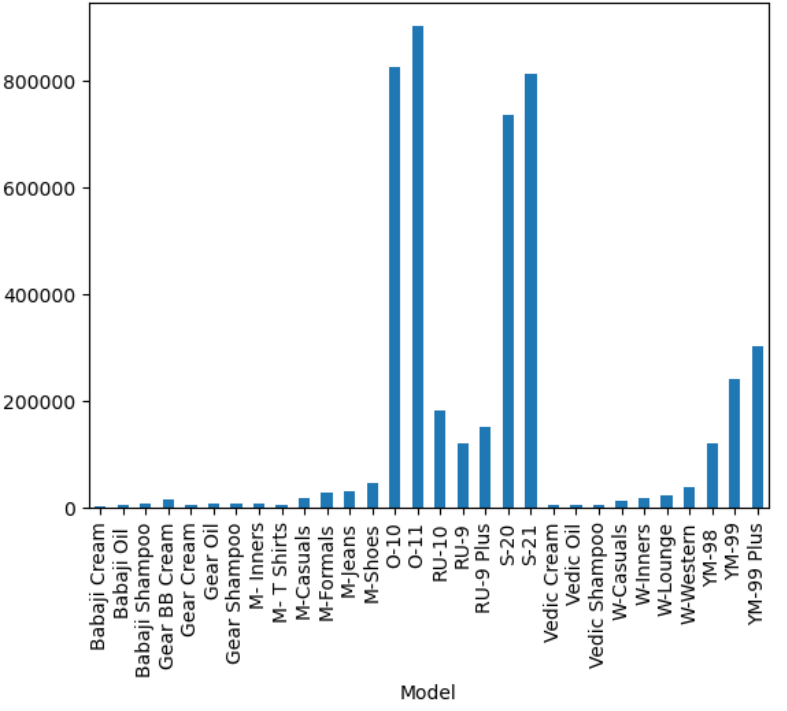


**Insights:**

* Highest frequency columns are F01,L01,M01
* Lowest frequency columns are F08,L10,M08.

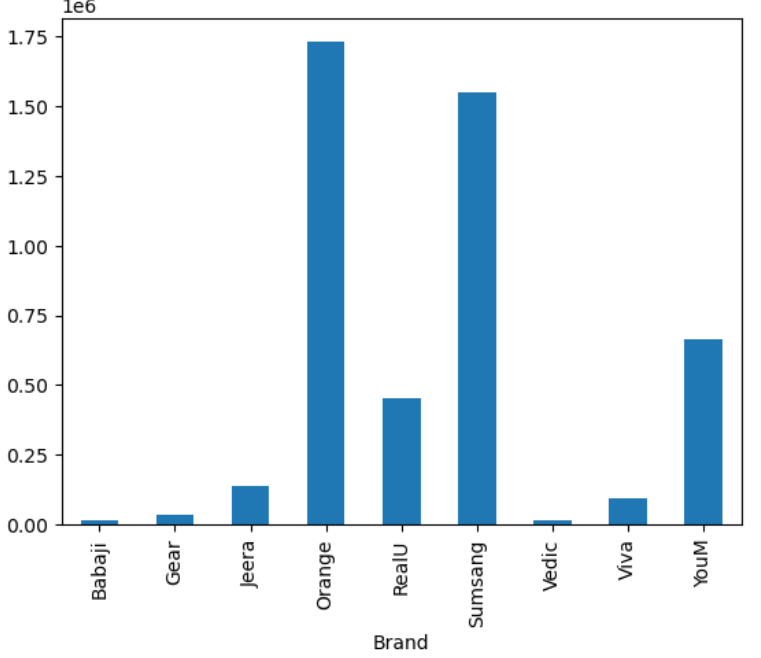
**Performing Bar Graph of “Avg Price” with categorical columns:**

* “Avg Price” Numerical Column with ‘Model’ Categorical Column.



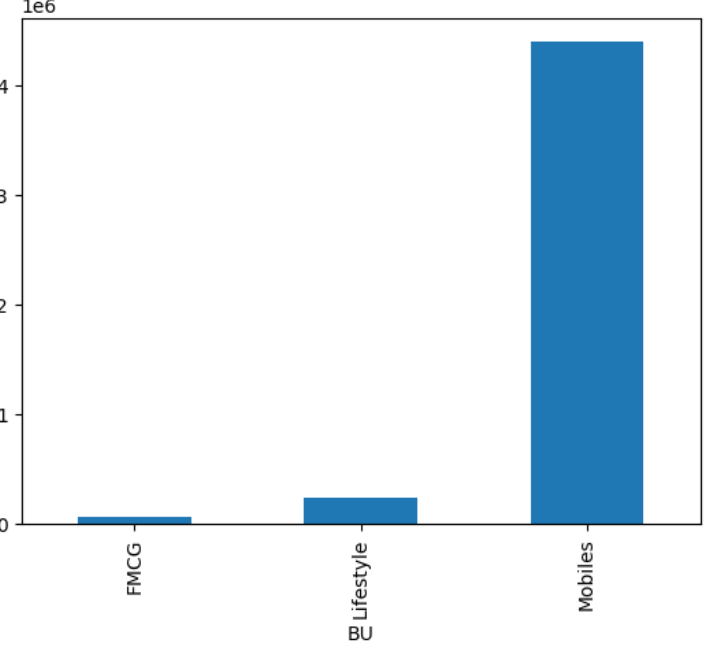
**Insights:**

* Highest frequency column is O11.
* Lowest frequency columns are Babaji Cream, Vedic Cream, Gear Cream, M-TShirts, Vedic Oil, Vedic Shampoo(approximately)…
* **Avg Price” Numerical Column with ‘Brand’ Categorical Column**.



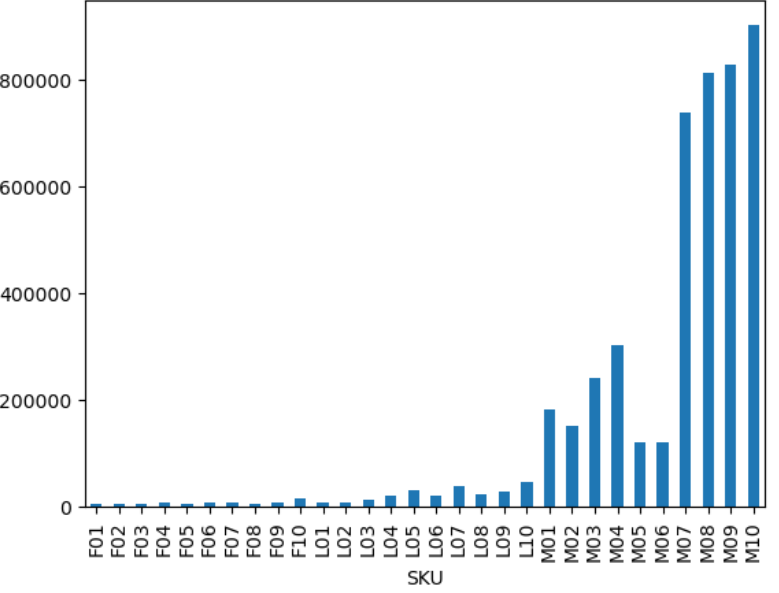
**Insights:**

* Highest frequency columns are Samsung and Orange.
* Lowest frequency columns are Babaji, Vedic and Gear
* **“Avg Price” Numerical Column with ‘BU’ Categorical Column**.



**Insights:**

* Highest frequency column is Mobiles.
* Lowest frequency columns can be observed from FMCG.
* **“Avg Price” Numerical Column with ‘SKU’ Categorical Column**.

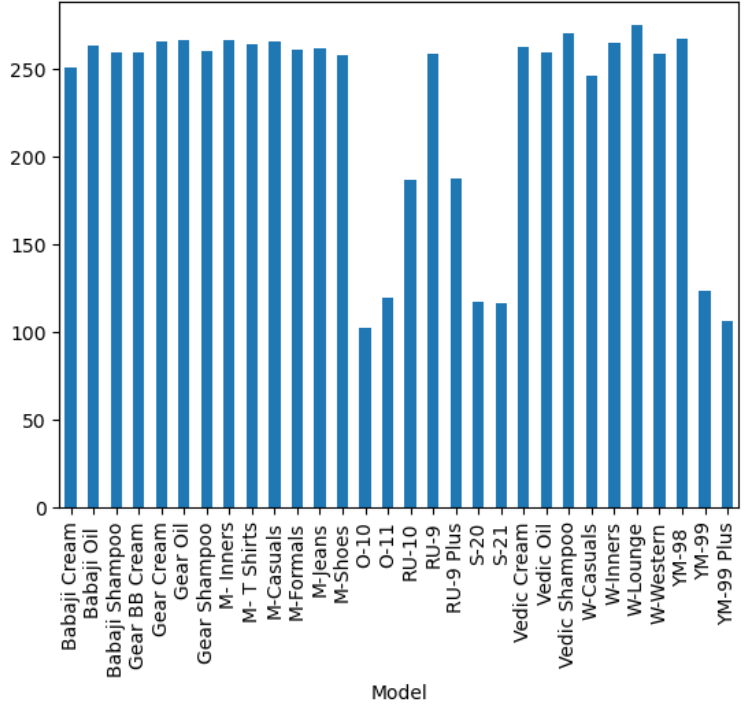


**Insights**:

* Highest frequency columns are MO10,MO9 and MO8.
* Lowest frequency columns can be observed from FO1 to L10(approximately).

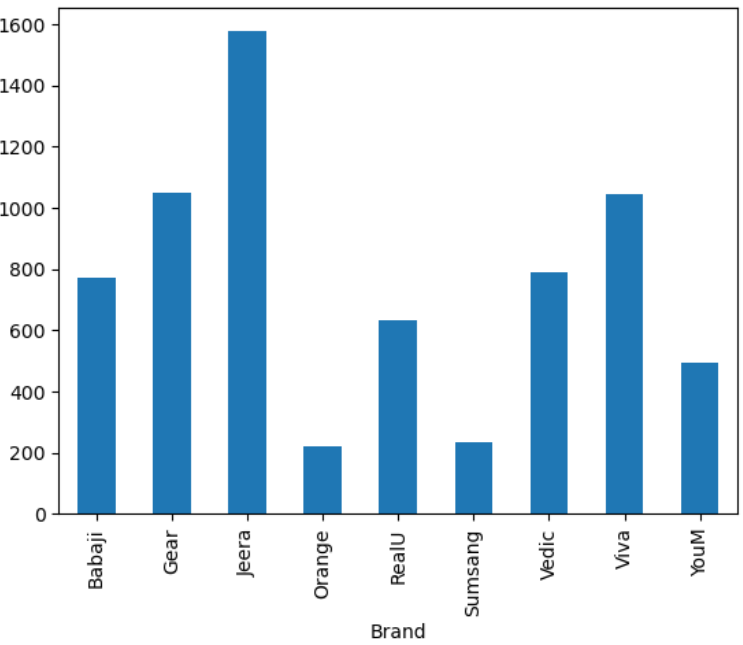
**Performing Bar Graph of “Discount Rate” with categorical columns:**

* “**Discount Rate” Numerical Column with ‘Model’ Categorical Column.**



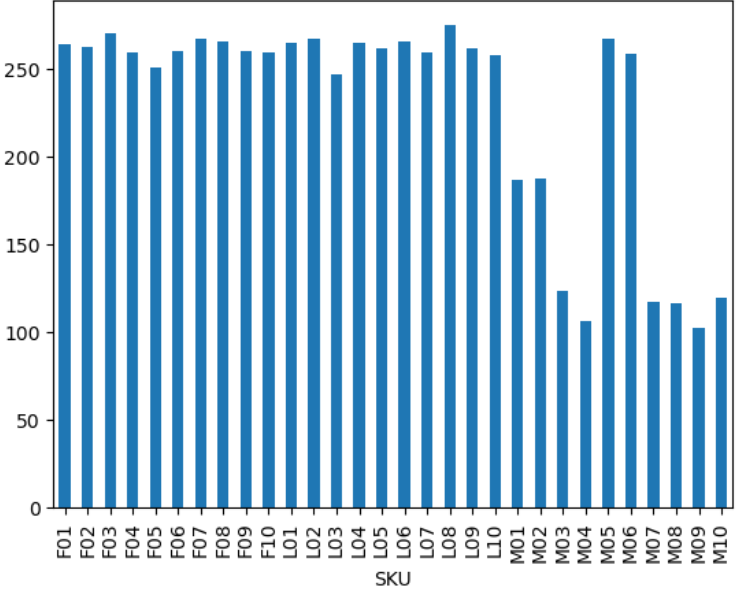
**Insights:**

* Highest frequency column is W-Lounge.
* Lowest frequency column is O-10.
* **Discount Rate” Numerical Column with ‘Brand’ Categorical Column.**



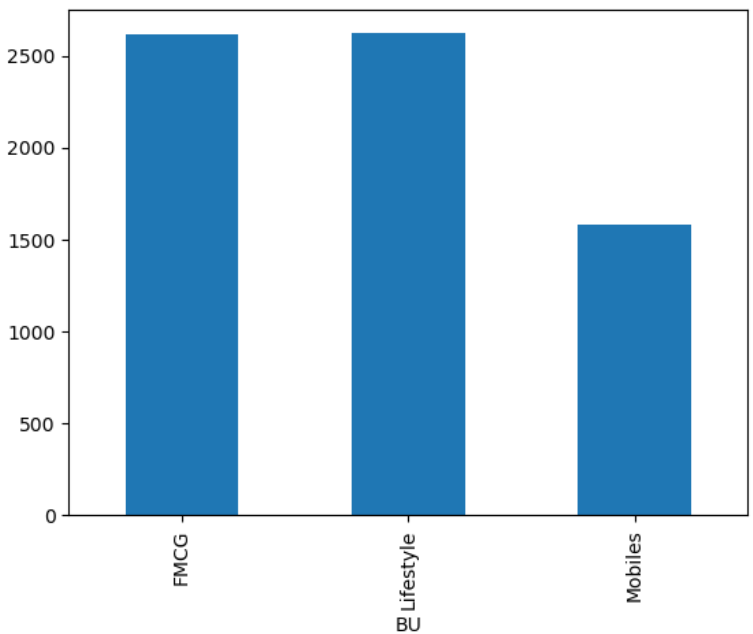
**Insights:**

* Highest frequency column is Jeera.
* Lowest frequency column is Orange and Samsung.
* **Discount Rate” Numerical Column with ‘SKU’ Categorical Column.**



**Insights:**

* Highest frequency column is LO8.
* Lowest frequency columns is MO4.
* **Discount Rate” Numerical Column with ‘BU’ Categorical Column**.

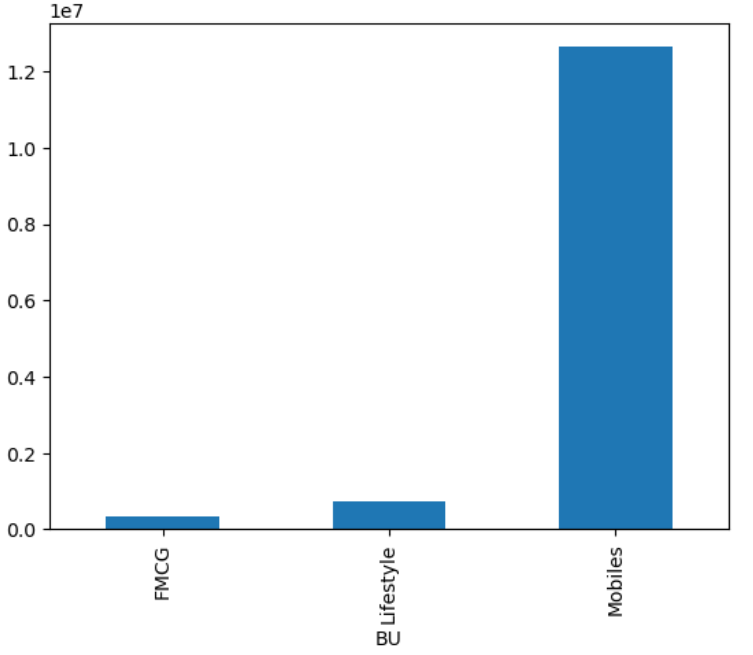


**Insights:**

* Highest frequency columns are FMCG and Life Style.
* Lowest frequency columns is Mobiles.

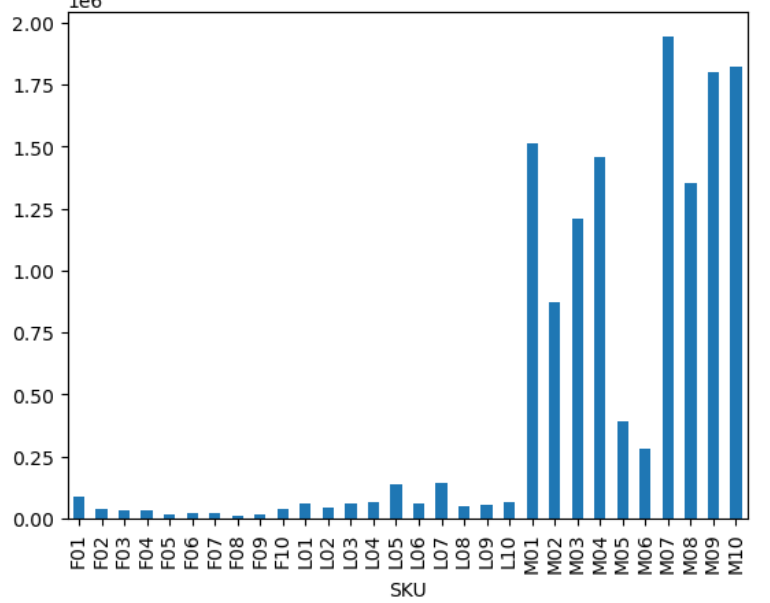
**Performing Bar Graph of “Net Sales Value” with categorical columns:**

* **“Net Sales Value” Numerical Column with ‘BU’ Categorical Column**.



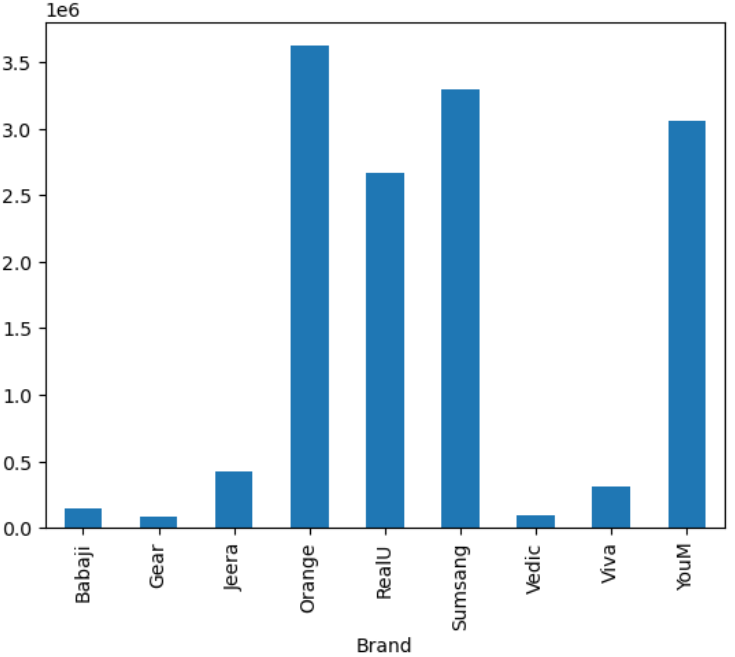
**Insights:**

* Highest frequency column is Mobiles.
* Lowest frequency column is FMCG.
* **“Net Sales Value” Numerical Column with ‘SKU’ Categorical Column**.



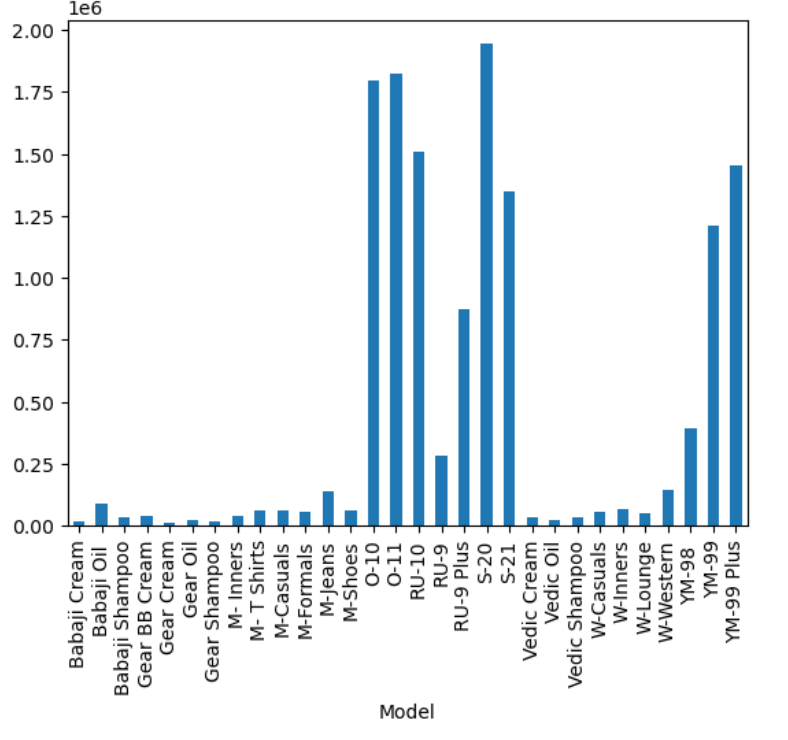
**Insights:**

* Highest frequency column is MO7.
* Lowest frequency columns are FO8,FO9,FO5(approximately).
* **“Net Sales Value” Numerical Column with ‘Brand’ Categorical Column.**



**Insights:**

* Highest frequency column is Orange.
* Lowest frequency column is Vedic and Gear(approx.)
* **“Net Sales Value” Numerical Column with ‘Model’ Categorical Column**.

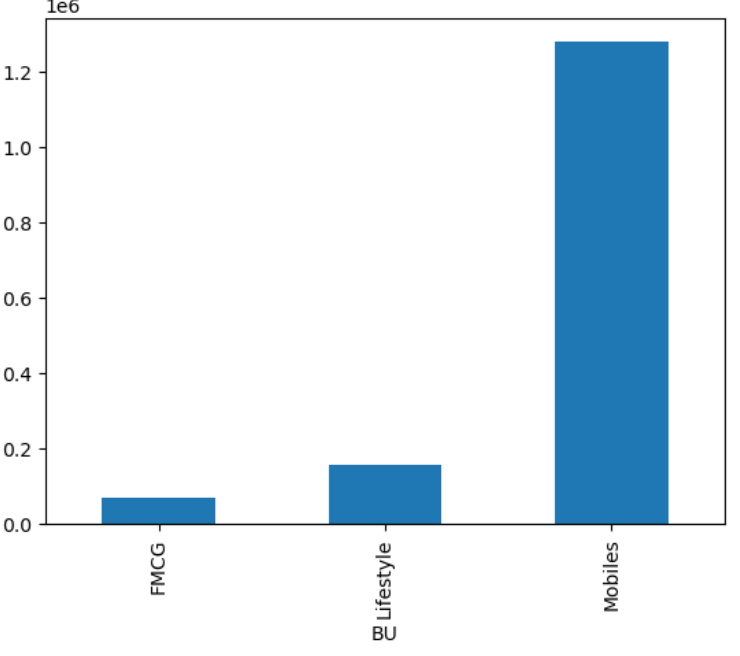


**Insights:**

* Highest frequency column is S-20.
* Lowest frequency columns are Babaji Cream, Gear Cream and Gear Shampoo(approx.).

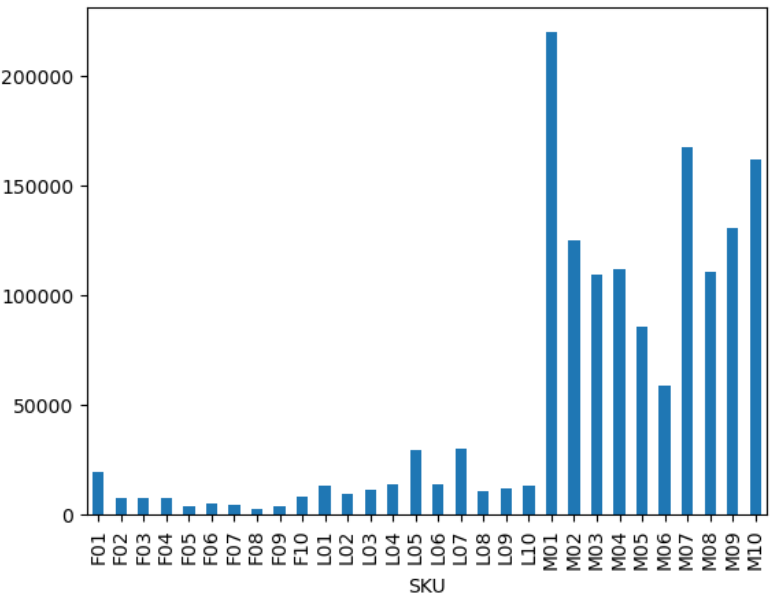
**Performing Bar Graph of “Discount Amount” with categorical columns:**

* **“Discount Amount” Numerical Column with ‘BU’ Categorical Column.**



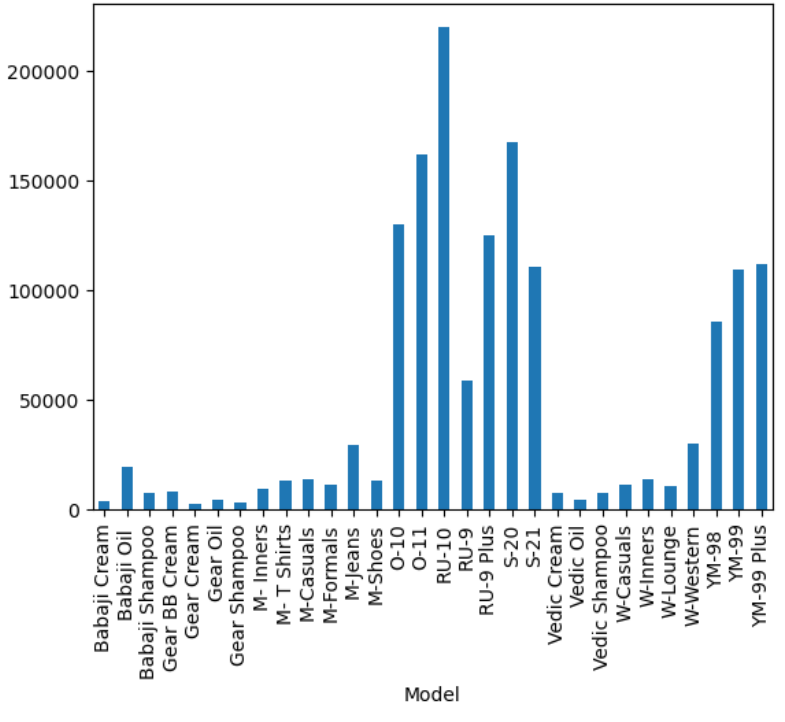
**Insights:**

* Highest frequency column is Mobiles.
* Lowest frequency column is FMCG.
* **“Discount Amount” Numerical Column with ‘SKU’ Categorical Column.**



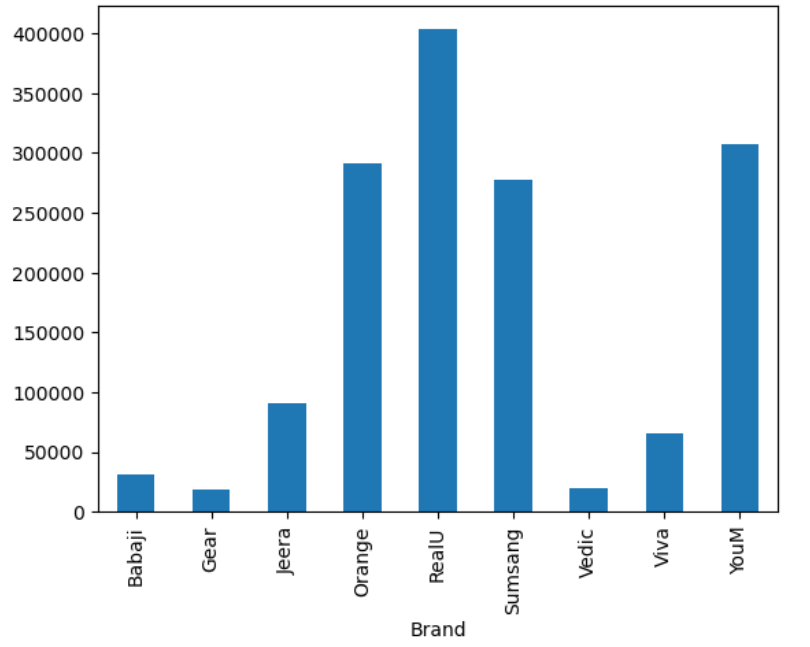
**Insights:**

* Highest frequency column is MO1.
* Lowest frequency column is FO8.
* **“Discount Amount” Numerical Column with ‘Model’ Categorical Column.**



**Insights:**

* Highest frequency column is RU-10.
* Lowest frequency column is Gear Cream, Gear Shampoo(approximately).
* **“Discount Amount” Numerical Column with ‘BU’ Categorical Column.**

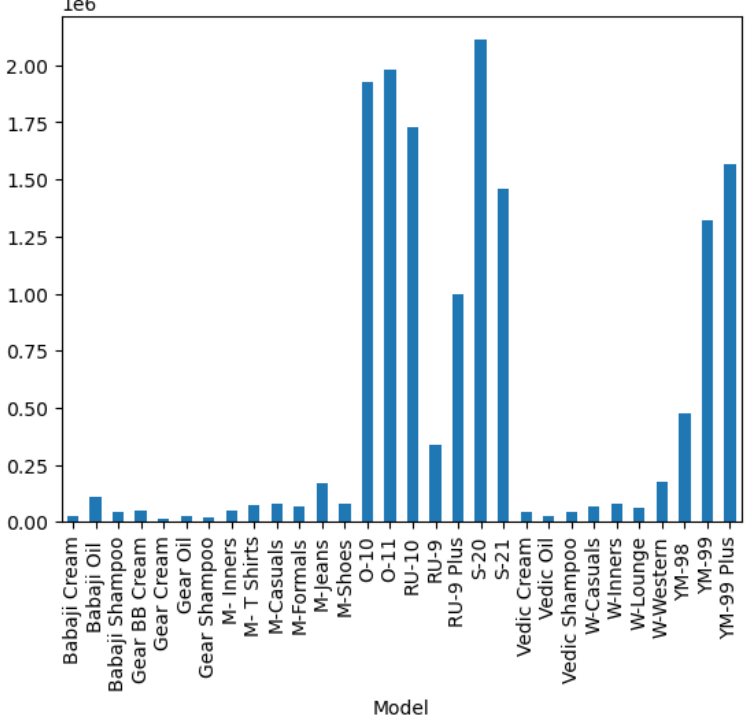


**Insights:**

* Highest frequency column is RealU.
* Lowest frequency column is Gear and Vedic(approximately).

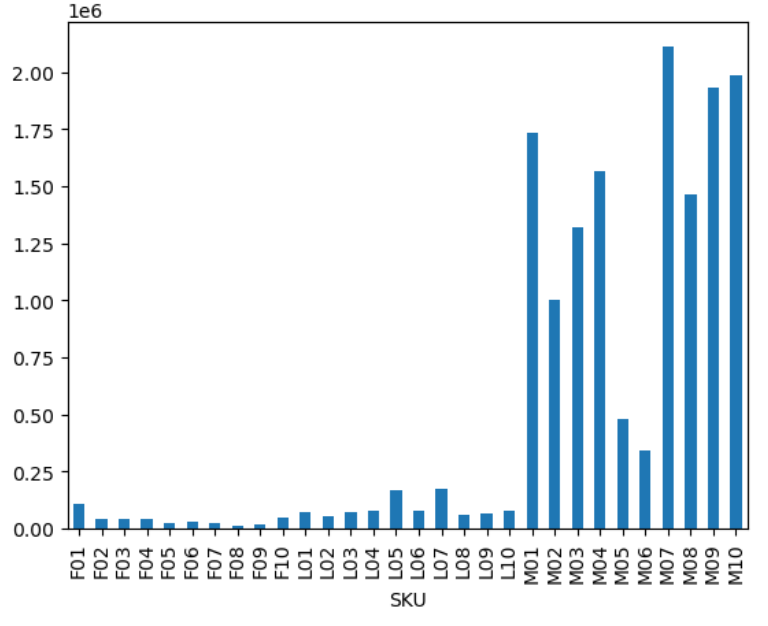
**Performing Bar Graph of “Total Sales Value” with categorical columns:**

* “**Total Sales Value” Numerical Column with ‘Model’ Categorical Column**.



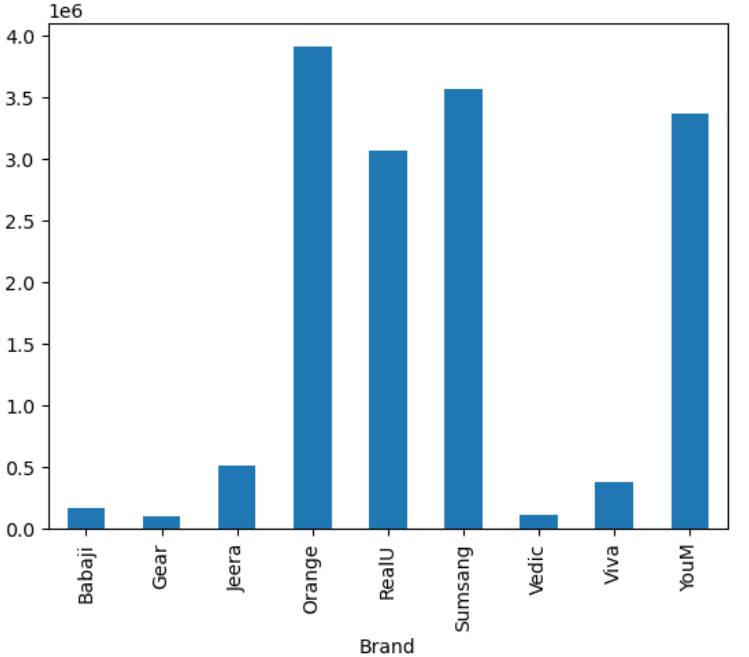
**Insights:**

* Highest frequency column is S-20.
* Lowest frequency columns are Gear Cream Gear Shampoo(approximately)
* **“Total Sales Value” Numerical Column with ‘SKU’ Categorical Column.**



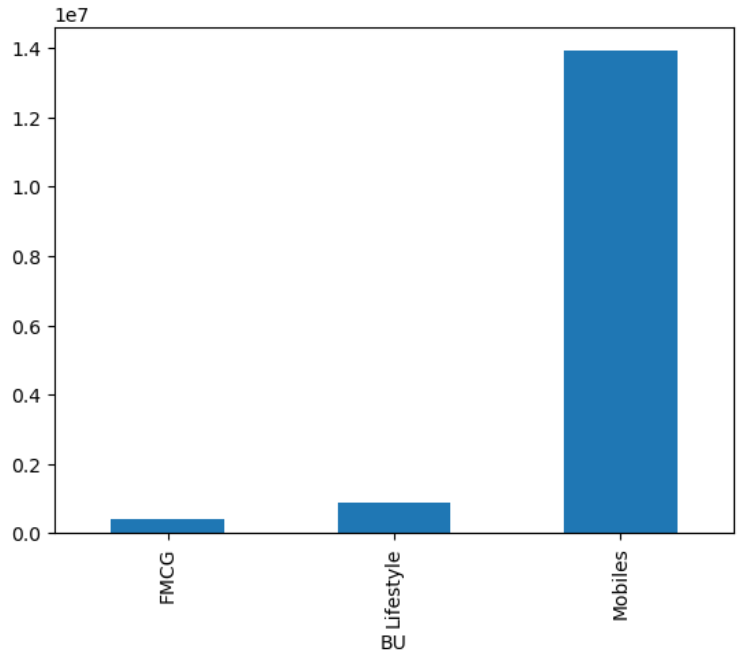
**Insights:**

* Highest frequency column is MO7.
* Lowest frequency column is FO8,FO9(approximately).
* **“Total Sales Value” Numerical Column with ‘Brand’ Categorical Column.**



**Insights:**

* Highest frequency column is Orange.
* Lowest frequency column is Vedic and Gear(approximately).
* **“Total Sales Value” Numerical Column with ‘BU’ Categorical Column.**



**Insights**:

* Highest frequency column is Models.
* Lowest frequency column is FMCG(approximately)

#### **Standardization of Numerical Variables:**

Z Standard normal variate function where it will transforms its data from its originally to around zero will in the interval of -3 to +3 sigma. Steps for standardization:

***Step 1:***

Calculating the Mean (μ) of the data.

***Step 2:***

Calculating the Standard Deviation (σ) of the data.

***Step 3:***

Transforming the Data using the formula:

z=x−μ/σ

​ where x is a data point, μ\mean is the mean of the dataset, and σ\sigma is the

standard deviation of the dataset.

In Z the average is 0 and Standard deviation is 1 always.

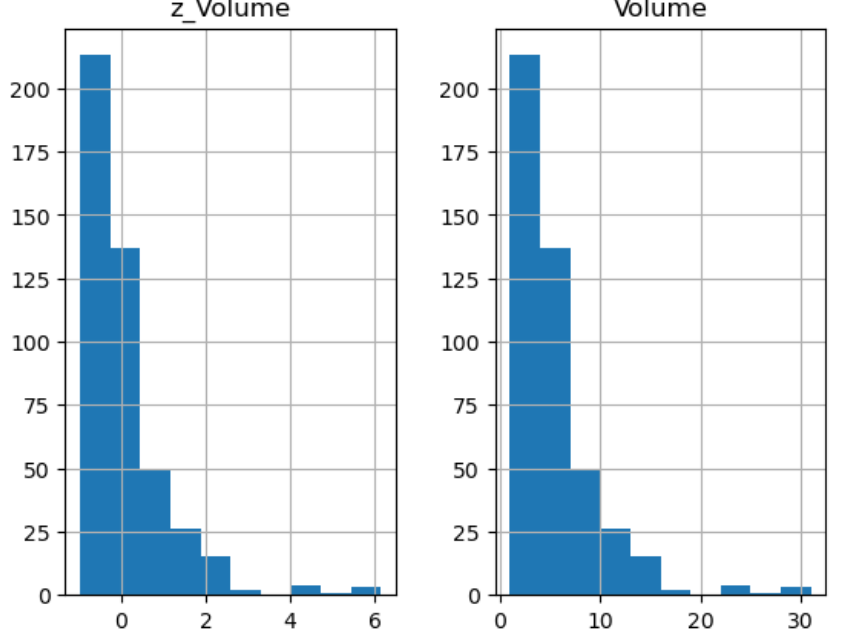
* Appoximately 68% of data lies between -1 sigma to +1 sigma from the center.
* Appoximately 95.5% of data lies between -2 sigma to +2 sigma from the center.
* Appoximately 99% of data lies between -3 sigma to +3 sigma from the center.

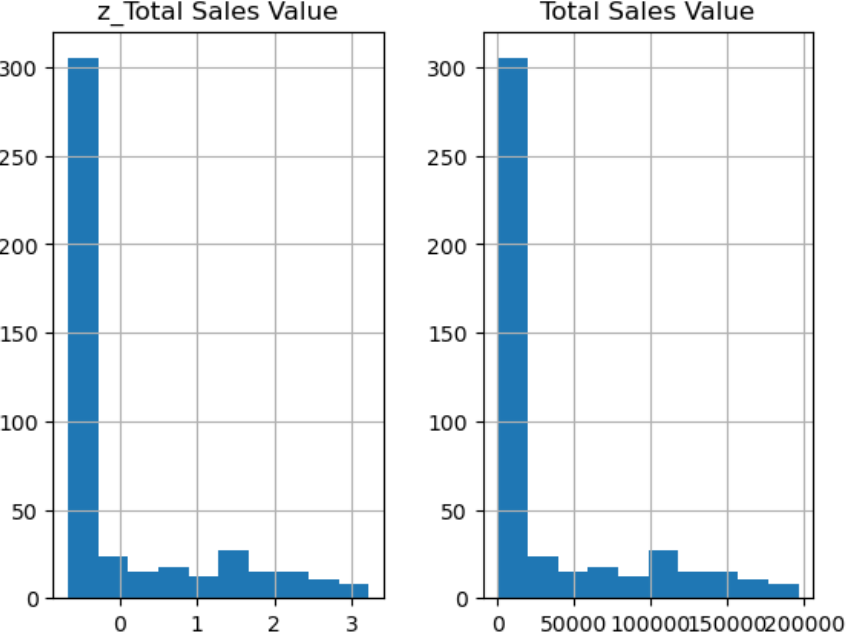
**Standardize the numerical columns using the formula: z=x-mu/sigma:**

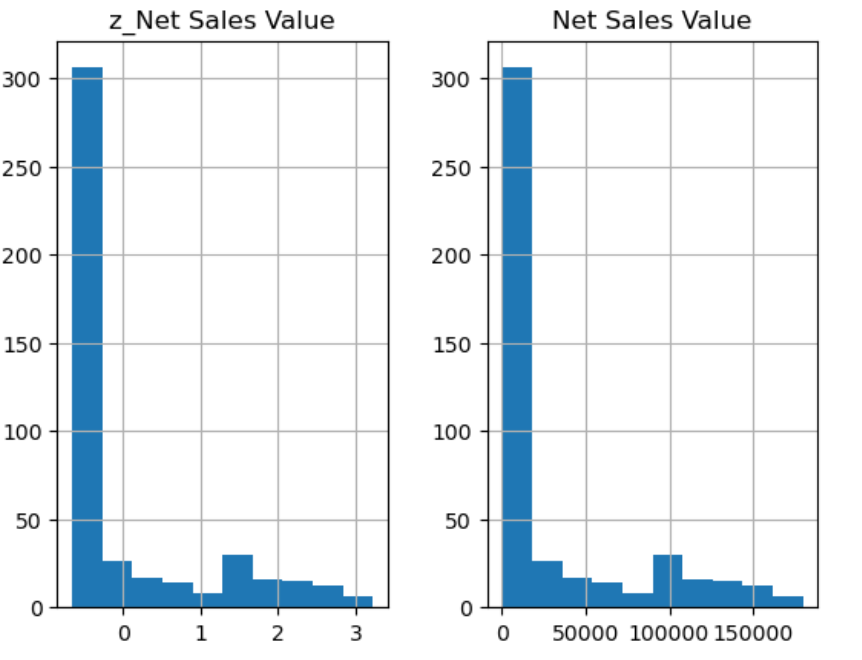
* df["Z\_Columnname"]=(df["Columnname"]df["Columnname"].mean())/df

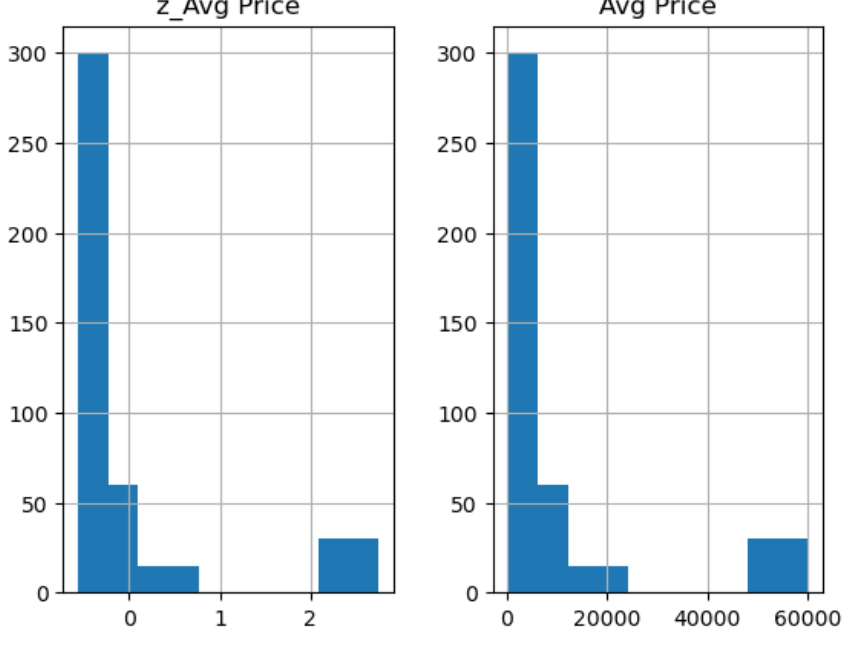
["Columnname "].std()

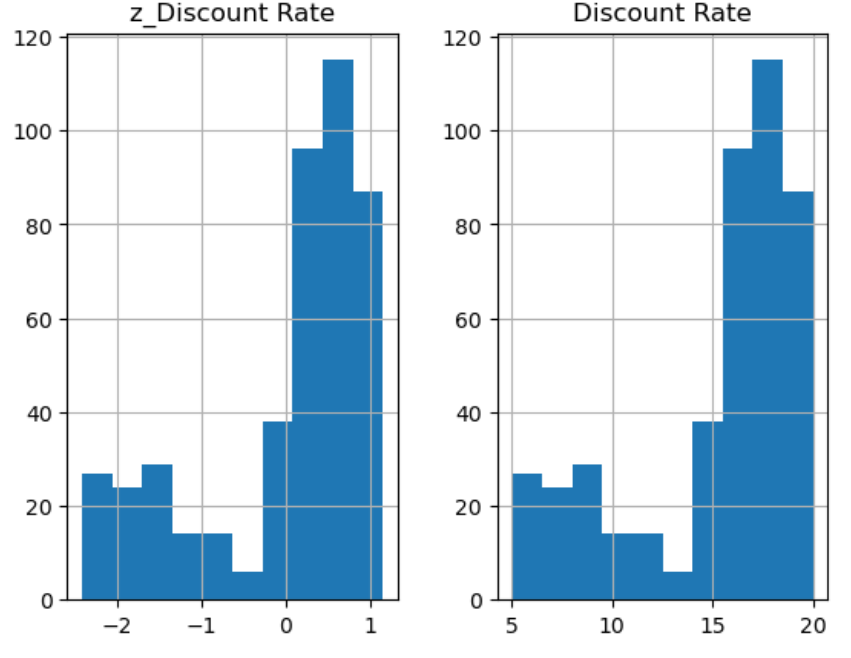
* df["Columnname "].mean().round(3)
* **Before and After comparisons of the Data distributions.**

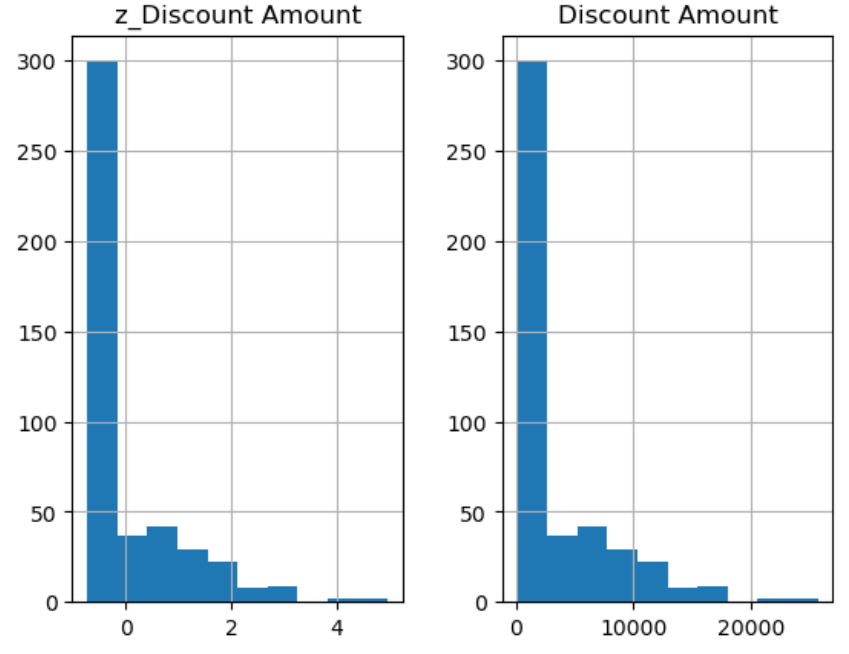












#### **Conversion of Categorical Data into Dummy Variables**

One-hot encoding is a crucial preprocessing step in machine learning that transforms categorical data into a suitable numeric format, avoids misleading ordinal assumptions, improves model performance, captures non-linearities.

**Apply one-hot encoding to the categorical columns**

* df\_cat = df[['Day','SKU','City','BU','Brand','Model']]
* df\_cat.head()
* from sklearn.preprocessing import OneHotEncoder
* OHE = OneHotEncoder()
* dummy = OHE.fit\_transform(df\_cat).toarray()
* dummy
* df\_cat\_full = pd.DataFrame(dummy)
* k1 = ['Friday','Monday','Saturday','Sunday','Tuesday','Thursday','Wednesday','F01','F02','F03','F04','F05','F06','F07','F08','F09','F10','L01','L02','L03','L04','L03','L05','L06','L07','L08','L09','L10','M01','M02','M03','M04','M05','M06','M07','M08','M09','M10','C','FMCG','Mobiles','Lifestyle','Babaji','Gear','Jeera','Orange','RealU','Sumsang','Vedic','Viva','YouM','Babaji Cream','Babaji Oil','Babaji Shampoo','Gear BB Cream','Gear Cream','Gear Oil','Gear Shampoo','M-Casuls','M-Formals','M-Jeans','M-T Shirts','M-Shoes','M-Inners','O-10','O-11','RU-9','RU-9 Plus','RU-10','RU-10','S-20','S-21','Vedic Cream','Vedic Oil','Vedic Shampoo','W-Casuals','W-Lounge','YM-98','YM-99','YM-99 Plus']
* df\_cat\_full.columns = k1
* df\_cat\_full.head()

**Displaying a portion of the transformed dataset:**

